2009-09-22

Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung

Jeffrey D. Gawronski
University of Massachusetts Medical School

Et al.

Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung

Jeffrey D. Gawronski[a], Sandy M. S. Wong[a], Georgia Giannoukos[b], Doyle V. Ward[b], and Brian J. Akerley[a,1]

[a]Department of Molecular Genetics and Microbiology, University of Massachusetts Medical School, 55 Lake Avenue North, S6-242, Worcester, MA 01655; and [b]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142

Edited by John J. Mekalanos, Harvard Medical School, Boston MA, and approved July 30, 2009 (received for review June 15, 2009)

Rapid genome-wide identification of genes required for infection would expedite studies of bacterial pathogens. We developed genome-scale "negative selection" technology that combines high-density transposon mutagenesis and massively parallel sequencing of transposon/chromosome junctions in a mutant library to identify mutants lost from the library after exposure to a selective condition of interest. This approach was applied to comprehensively identify *Haemophilus influenzae* genes required to delay bacterial clearance in a murine pulmonary model. Mutations in 136 genes resulted in defects in vivo, and quantitative estimates of fitness generated by this technique were in agreement with independent validation experiments using individual mutant strains. Genes required in the lung included those with characterized functions in other models of *H. influenzae* pathogenesis and genes not previously implicated in infection. Genes implicated in vivo have reported or potential roles in survival during nutrient limitation, oxidative stress, and exposure to antimicrobial membrane perturbations, suggesting that these conditions are encountered by *H. influenzae* during pulmonary infection. The results demonstrate an efficient means to identify genes required for bacterial survival in experimental models of pathogenesis, and this approach should function similarly well in selections conducted in vitro and in vivo with any organism amenable to insertional mutagenesis.

Illumina | mariner | mutagenesis | pathogenesis | transposon

Whole-genome analytic techniques have been developed to identify bacterial genes essential for growth or survival in vitro or during infection of model hosts. The most direct of these approaches can be classified as "negative selection" strategies, in which large pools of diverse mutants are analyzed to identify mutations that reduce fitness under a particular condition. "Signature-tagged mutagenesis" utilizes DNA arrays representing unique hybridization tags that are introduced into each mutant within a library of strains to be evaluated (1). The "transposon-site hybridization" and "microarray tracking of transposon mutants" methods use microarrays displaying each gene of the target organism to monitor the relative abundance of transposon insertions in these genes under varied selection conditions (2–4). Each of these methods has been effectively used to identify virulence genes in diverse bacteria. For many pathogens, however, generation of large banks of uniquely tagged mutants is impractical and whole-genome microarrays may be unavailable, particularly for newly recognized organisms or genetically diverse species. In both microarray-based methods, hybridization is used to detect the abundance of a given mutation within the library of mutants. Therefore, quantification is limited by background hybridization levels and the dynamic range of signal detection. A method that generates an output that allows precise noise filtering and a broad dynamic range would represent a significant advancement of the negative selection strategy.

In this study we report a technique termed "high-throughput insertion tracking by deep sequencing" (HITS) that uses a whole-genome transposon mutant bank in combination with massively parallel sequencing to efficiently analyze bacterial genes involved in pathogenesis. HITS allowed analysis of genes required by *Haemophilus influenzae* to resist clearance from the lung, a site colonized during pneumonia and chronic obstructive pulmonary disease (5, 6). Because deep sequencing is used for detection, background signal is easily identified and removed during data analysis, and the dynamic range of detection is limited only by the number of sequencing reads, which can be readily increased. The results highlight the utility of HITS in systematic discovery and analysis of virulence genes required in environments encountered by bacteria during pathogenesis.

Results

Overview of the HITS Technique. HITS is outlined schematically as two steps in Fig. 1 *A* and *B*. The first step involves fragmentation and ligation of adapters to sheared genomic DNA prepared from a high-density mutant bank carrying random transposon insertion mutations. In this study, mutagenesis was performed with a minitransposon derived from the *Himar1-mariner* transposon, which inserts efficiently in the genomes of *H. influenzae* and other bacteria, with only the dinucleotide TA as the apparent insertion site specificity (7–9). Selective amplification of transposon/chromosome junction regions is performed by PCR, and the resulting amplicons are purified by affinity capture. Sequencing is performed *en masse* on the Illumina next-generation sequencing platform. The second step identifies the genomic location of each transposon insertion site within the bank by mapping chromosomal sequences adjacent to inverted terminal repeats of the transposon to the reference genome. The fitness of insertion mutants containing disruptions in a given gene is reflected in both the relative number of insertion sites detected within the gene and the number of times each site is detected by the sequence analysis.

Generation of a Mutant Bank Selected for Growth or Survival in Vivo in the Murine Lung Model. The mechanisms that allow *H. influenzae* to persist in the lung are not well understood. Mouse pulmonary infection provides a well-established model for investigation of mechanisms used by *H. influenzae* and other

Author contributions: J.D.G., S.M.S.W., and B.J.A. designed research; J.D.G., S.M.S.W., and G.G. performed research; J.D.G., G.G., and D.V.W. contributed new reagents/analytic tools; J.D.G., S.M.S.W., and B.J.A. analyzed data; and J.D.G., S.M.S.W., and B.J.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Draft genome sequences for *H. influenzae* strains have been submitted to the National Center for Biotechnology Information: Rd BA042 (RdAW) (accession no. ACSN01000000) and NT127 (accession no. ACSL01000000).

[1]To whom correspondence should be addressed. E-mail: brian.akerley@umassmed.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0906627106/DCSupplemental.

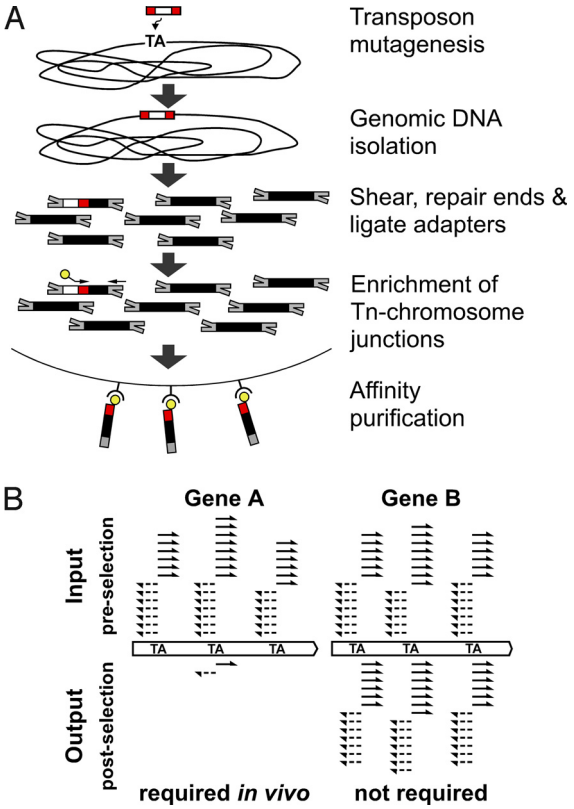

Fig. 1. HITS and comparison of selected libraries. (*A*) HITS sample preparation and enrichment of transposon/chromosome junctions. After transposon mutagenesis, chromosomal DNA is purified from *H. influenzae* mutant library. Red, ITRs of *himar1* transposon; white, contents of transposon, including the kanamycin resistance gene. Illumina oligonucleotide adapters (gray) are ligated to sheared genomic DNA. Fragments of the transposon/chromosome junctions are enriched via PCR using transposon- and adapter-specific primers. The biotinylated transposon-specific primer (yellow) anneals to the ITRs of the transposon and includes the Illumina sequencing primer site. The adapter-specific primer anneals to only 1 oligonucleotide of the partially complementary adapter. Enriched fragments are collected using streptavidin-coated paramagnetic beads. After washing, single-stranded DNAs are eluted from the beads and used for cluster formation on Illumina flow cells. (*B*) Comparison of lung-selected output library to input library. After sequencing, reads are mapped to the reference genome (solid arrows, plus strand; dashed arrows, minus strand) to identify the transposon insertion sites. The number of insertion sites detected per gene and the number of sequencing reads per site are used to determine the relative abundance of the mutant within the library before and after selection. The examples depict insertion patterns at TA sites in hypothetical gene A, in which insertion mutations confer attenuated growth or survival during infection, and gene B that is not required for growth in vitro or in vivo. Insertions in genes that are essential for growth on rich culture media are absent in the input library and are not detected by HITS.

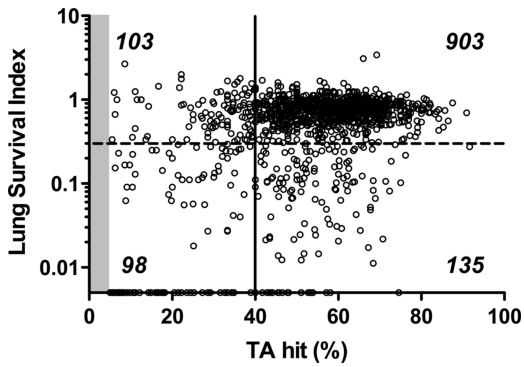

Fig. 2. Comparison of transposon insertions in the mutant library before and after selection in the lung model. Insertion sites in the 5′ 5–80% protein coding sequence of the gene and reads associated with these sites were considered for fitness analysis. The saturation of transposon insertions within 1,239 genes in the input library is shown on the *x* axis. Saturation was calculated as the percentage of sites within a gene sustaining transposon insertions to the total number of possible of insertion sites (TA dinucleotides). The lung survival index (s.i.) is represented on the *y* axis as the number of reads mapped to a gene in the output library divided by the reads identified in the input library (points on the *x* axis represent s.i. values of zero). Essential genes, those sustaining insertions in <5% of possible sites, are not shown (shaded); the majority of these sustained no insertions, and the remaining 25% averaged 1 insertion per gene. The threshold for an inferred in vitro growth defect (solid line) was set at a saturation of 40% of the possible TA insertion sites within a gene. The threshold for in vivo attenuation (dashed line) was set at a lung s.i. of <0.30. Numbers of genes falling within each quadrant are indicated.

bacteria to persist and resist host defenses during lung pathogenesis (10–12), yet there have been no comprehensive studies to identify *H. influenzae* genes needed at this site. To evaluate the utility of HITS for virulence gene identification using the mouse lung model of infection, we inoculated 5 mice with 10^7 cfu of a ≈75,000 member insertion mutant library of *H. influenzae* generated with a *Himar1 mariner*–derived minitransposon. At 24 h after inoculation an average of 9.2×10^5 cfu were recovered from the lungs of each mouse. Chromosomal DNA was isolated for analysis from both the inoculum and from the ex vivo bacterial populations. The numbers of cfu in the inoculum and recovered from mice suggested that the mutant library was likely to be sufficiently represented in both populations and that mutants had been subjected to in vivo selection.

Analysis of Genomic Mutant Banks by HITS and Application to Genome-Wide Identification of *H. influenzae* Genes Required in the Lung. We conducted the HITS procedure (Fig. 1) on the input library and mapped the insertions to their chromosomal positions (Fig. S1). Insertions were evenly distributed around the chromosome, and ≈44% of the 131,960 total possible chromosomal TA target sites for *mariner* were found to have sustained insertions in this library. Before passage in vivo, a total of 534,567 sequencing reads mapped to nonrepetitive chromosomal regions immediately flanking 55,935 unique sites, with 44,270 in predicted protein coding genes and 11,665 in intergenic regions or structural RNAs. Of 1,657 annotated genes, no insertions were detected within 268 genes and 90 sustained insertions in <5% of their possible TA insertion sites, implicating at least 358 genes as essential for growth or viability on laboratory medium in vitro. Twenty-five genes with <8 possible TA insertion sites were excluded from analysis on the basis of an estimated probability of 0.05 that they could fail to sustain insertions owing to chance at this level of transposon insertion density in the library. Thirty-five genes could not be analyzed because they either contained extensive repetitive sequences or were duplicated in the genome. After subtraction of essential genes, genes containing repetitive sequence, or genes with very few possible insertion sites, there were 1,239 genes that could be analyzed.

For fitness analysis of mutants after in vivo selection, we considered the number of transposon insertions detected in the first 5–80% of each gene, the region in which insertions are expected to abrogate gene function. To exclude genes in which insertion mutations led to potential in vitro growth defects, we set a threshold requiring that candidate virulence genes sustain insertions in at least 40% of their possible *mariner* transposon target sites before in vivo selection, and 201 genes sustained densities of insertions below this threshold (Fig. 2 and Table S1). In the 1,038 annotated protein-coding genes that were dispensable in vitro, an average of 287 sequencing reads detecting insertions in the 5′ 5–80% region of each gene was observed

MICROBIOLOGY

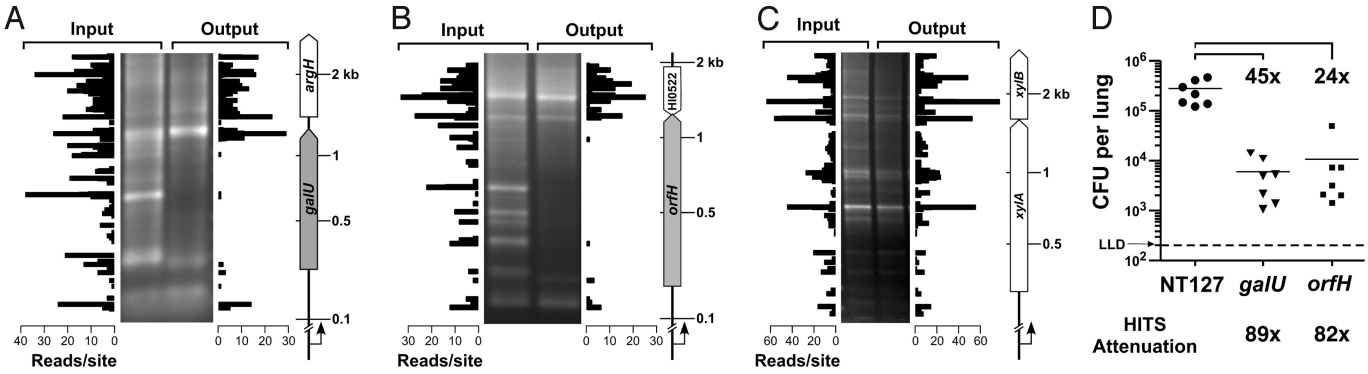

Fig. 3. Comparison of HITS analysis, genetic footprinting, and single-strain infections. Genetic footprinting of input and output libraries for (*A*) *galU*, encoding UDP-glucose pyrophosphorylase, (*B*) *orfH*, encoding heptosyltransferase III, and (*C*) *xylA*, encoding xylose isomerase, are shown in the gel images. PCR analysis was conducted using the transposon-specific primer marout and chromosomal primers galU_F, orfH_F, or xylA_F that anneal 278 bp, 202 bp, and 279 bp upstream of the respective genes. In the plots, HITS data correspond to regions analyzed by footprinting. (*D*) *H. influenzae* NTHi wild-type (NT127) and deletion mutants of *galU* and *orfH* recovered from lungs of C57BL/6 mice (7 mice per strain) 24 h after intranasal inoculation with each strain. Bars represent the mean cfu per lung. Comparisons between wild-type and mutants were statistically significant via one-way ANOVA with Tukey's multiple comparison test ($P < 0.001$). LLD, lower limit of detection. Fold differences in mean cfu recovered for NT127 wild-type strain vs. the *galU* or *orfH* mutants in individual mutant infections (brackets) are compared with HITS results (below the chart). HITS survival indices were 0.011 for *galU* and 0.012 for *orfH*, corresponding to in vivo attenuations (calculated as the reciprocal of the s.i.) of 89-fold and 82-fold, respectively. In *A–C*, genome coordinates of transposon insertion sites detected via HITS analysis were reoriented with respect to the chromosomal primer positions used in footprinting. The *y* axis was modeled to the migration of the molecular weight (MW) standard of footprinting gels using nonlinear regression, and the *x* axis represents the number of sequencing reads mapped to insertion sites. The scale of the MW standards on the right of each panel applies to both the genetic footprints and the HITS analysis plots. White, nonessential genes; gray, genes required for growth or survival in the lung.

(Fig. S2). HITS analysis of these genes was quite reproducible. When two preparations of genomic DNA from the transposon mutant bank were independently analyzed, the number of insertions detected in each gene was similar, with the majority (82%) of genes having <20% variation in insertion density between samples (Fig. S2). Therefore, both the complexity of the transposon bank and the detection of mutations by sequencing seemed to be sufficiently saturating for reproducible analysis of the relative abundance of mutants in the library.

To identify genes required during infection, we analyzed the relative number of insertions in each gene in the output library obtained after lung infection vs. insertions in the library before in vivo selection. The results are shown graphically in Fig. 2, and the complete data are listed in Table S1. A total of 903 genes had similar numbers of insertions before and after selection in the lung, indicating that they were not required in this model. This large number of genes with insertion patterns in the output library that were similar to those in the input library indicated that significant stochastic loss of mutations had not occurred in the infection model. The 135 genes that sustained insertions in >40% of their possible TA insertion sites and in which the number of insertions decreased by at least 3.3-fold after selection in the lung were considered candidate virulence genes (Table S2). Representative insertion patterns for genes detected as being required during infection (*galU* and *orfH*) vs. those that are not required in vivo (*xylA*) are shown in Fig. S3. In summary, HITS implicated 8.1% of the 1,657 annotated genes in the genome in survival or growth of *H. influenzae* in the lung.

Genetic footprinting provides a means for analyzing insertions in discrete genes to verify results obtained with HITS. Genetic footprinting uses PCR with a specific chromosomal primer paired with a transposon primer for physical mapping of insertions to the chromosome in a bank of mutants (13). For a given gene, PCR results in a set of products varying in size that correspond to the distance between the chromosome-specific primer and each transposon mutation within that gene. Specificity is further assured by conducting the procedure with a primer 5′ of the gene and independently with a primer 3′ of the gene. For this validation we chose genes of LPS biosynthesis (*opsX*, *rfaF*, *orfH*, and *galU*) in which mutations resulted in pronounced attenuation relative to wild-type according to HITS data (Table S2). *H. influenzae* produces a short chain carbohydrate on its LPS (also called lipooligosaccharide, LOS) and lacks the repeating O-antigen carbohydrate typical of some bacterial LPS. The LOS of *H. influenzae* consists of a conserved "inner core" usually composed of 3 heptose residues and an "outer core" composed of variable-length oligosaccharide extensions from the heptose residues. The *opsX*, *rfaF*, and *orfH* genes encode heptosyltransferases I, II, and III, respectively, and generate the chain of 3 heptose residues initiating at a single 3-deoxy-D-*manno*-octulosonic acid, which is attached to lipid A (14–16). The *galU* gene encodes a UDP-glucose pyrophosphorylase that catalyzes the UTP-dependent conversion of D-glucose-1-phosphate into UDP-glucose, the activated form of the sugar required for biosynthesis of various carbohydrates, and *galU* is required for addition of glucose and galactose residues to the LPS of diverse pathogenic bacteria (17–19).

Representative genetic footprinting results are shown for *galU* and *orfH* and compared with insertion patterns detected by HITS in Fig. 3 *A* and *B*. The decrease in insertion mutations detected in these genes after in vivo passage of the bank provided verification of selection against mutants with disruptions in these genes, and band intensities on genetic footprinting gels were in good agreement with the abundance of insertions at each site as detected by HITS. Genetic footprinting also detected in vivo attenuation of mutants with insertions in *opsX*, *rfaF*, and *galE*, and similar results were obtained in reactions with primers positioned either 5′ or 3′ of each gene (Fig. S4). In contrast, *xylA*, a gene of D-xylose metabolism that is not required for bacteremia (20), exhibited similar mutational profiles in both the input and output banks (Fig. 3*C*), indicating that it is dispensable in the lung model. Therefore, these results provided a verification of HITS results by an independent method, identifying virulence factors previously implicated in bacteremia.

To assess whether genes identified by HITS as being required in vivo are also required in single-strain infections, we generated nonpolar mutations removing the complete coding sequences of *galU* or *orfH*, genes implicated as being required in the lung model by HITS. To address mutant phenotypes with a recent clinical isolate, mutations were constructed in the nontypeable

H. influenzae strain, NT127 (21). In agreement with HITS and genetic footprinting results, both mutants were attenuated for survival in the lung. Moreover, the degree of attenuation calculated by HITS falls within the variation in fold difference observed between single-strain infections of individual mice. (Fig. 3*D*). The *galU* and *orfH* genes were previously shown to be essential for survival of *H. influenzae* in bloodstream models of infection (22, 23). A requirement for these genes in the lung supports the view that *H. influenzae* utilizes structures of the LPS inner and outer core in virulence strategies to combat clearance mechanisms of the host found in both of these environments.

Discussion

The genes implicated in bacterial growth or survival in the lung were functionally diverse, although several general categories were notable (Table S3). On the basis of Clusters of Orthologous Groups (COG) classifications, categories that were overrepresented in the attenuated gene set relative to their representation in the genome overall were "cell wall/membrane/envelope biosynthesis," "amino acid transport and metabolism," and "nucleotide transport and metabolism" (Tables S2 and S3). The genes identified provided insight into the selection conditions encountered by *H. influenzae* in the lung model.

Components of the bacterial cell surface are frequently the most direct participants in host–pathogen interactions. A major class of genes related to the cell envelope that was identified as markedly attenuated in vivo consisted of genes of LPS synthesis. LPS is essential in models of *H. influenzae* pathogenesis in the middle ear and blood and contributes to numerous aspects of NTHi infection, including evasion of complement and antimicrobial peptides (24–26). Genes needed for extension of the LPS inner-core structures (*opsX*, *rfaF*, and *orfH*) were required in vivo in the lung [survival index (s.i.) $\leq$0.012] (Table S2), in agreement with the requirement for these genes for bacteremia (22, 23). Genes required for precursor production for LPS carbohydrate outer-core hexose extensions (*galU* and *galE*) were also required (s.i. $\leq$ 0.025), suggesting that unmodified inner-core LPS results in enhanced clearance of *H. influenzae* from the lung. Genes required for hexose extensions from the first heptose, *lgtF* (s.i. = 0.152), or the terminal heptose of the inner core, *lpsA* (s.i. = 0.258), were partially required (16, 27), and a trend of moderate attenuation ($\approx$1.5-fold) was also observed in single-strain infections with an *lpsA* mutant (Fig. S5). Distal modifications of the LPS outer-core structure mediated by genes such as *lic3A*, which adds sialic acid or the *lic1* locus responsible for addition of phosphorylcholine seemed to be nonessential in vivo in these experiments. The *lic1D* gene was previously implicated in the lung model at a late time during infection, but not at 24 h (28), and it is possible that other distal modifications also are more important at later times.

Numerous genes involved in transport of proteins or other substrates were implicated in the lung model, including the complete twin-arginine translocation system (*tatA*, *tatB*, and *tatC*), which translocates folded proteins that lack Sec-dependent signal sequences across the plasma membrane and contributes to virulence in multiple pathogens (29). An intriguing set of genes with recently predicted functions in maintenance of outer-membrane lipid asymmetry (30) was implicated in pathogenesis in the lung. These genes included *vacJ* and a set of 5 genes annotated as "hypothetical genes" that are putative orthologs of an ABC transport system of *Escherichia coli* encoded by the *mlaA and mlaBCDEF* loci (for clarity, *E. coli* names of these genes are noted in Table S2). Orthologs of these genes were implicated in virulence of enteroinvasive *E. coli*, *Shigella flexneri*, and *Burkholderia pseudomallei* (31–33). The *mla* gene orthologs were required late during the intracellular life cycle for escape from the phagocytic vacuole (31). A role for the *mla* genes in both intracellular pathogens and *H. influenzae* suggests that *H. influenzae* may encounter membrane-damaging host defenses, such as cationic peptides or stress conditions in the lung, that are similar to those found in the phagocytic vacuole.

Additional overrepresented COG groups included genes involved in nutrient acquisition and interrelated adaptations to physiologic stress. Pathways of amino acid metabolism were required in the lung and included enzymes for synthesis or interconversion of methionine, asparagine, aspartate, serine, tryptophan, and branched-chain amino acids. Consistent with amino acid limitation, genes predicted to encode regulators involved in the stringent response were implicated in vivo, including RelA, which synthesizes (p)ppGpp in response to amino acid starvation (34), DksA, which modulates rRNA expression in response to (p)ppGpp (35), and Lon protease, which is activated by polyphosphate generated from (p)ppGpp (36) and has been implicated in proteolytic control of virulence factors (reviewed in ref. 37). Genes of nucleotide uptake and metabolism included those required for synthesis of purines and pyrimidines, in addition to genes involved in NAD uptake, *nadN* and *hel*. These genes mediate sequential conversion of NAD to NMN and nicotinamide riboside for uptake of this nucleotide that *H. influenzae* is unable to synthesize de novo (38). The complete set of genes for phosphate uptake (*pstS*, *pstB*, *pstA*, and *pstC*) was implicated in pathogenesis, as was the gene predicted to encode PhoB, a conserved response regulator protein that becomes active under low-phosphate conditions and controls diverse virulence functions in bacterial pathogens (reviewed in ref. 39). PhoR, a sensor kinase that activates PhoB, was not implicated in the lung. In other species, PhoB can be activated by "cross-talk" with other signaling systems independently of PhoR (40), and therefore *H. influenzae* PhoB may be required for responses to alternative signals. Resistance to oxidative stress is important for many pathogens. Genes involved in adaptations to oxidative stress conditions were identified, including *pgdX*, encoding a glutathione-dependent peroxidase (41), *oxyR*, which regulates genes critical for oxidative stress resistance, including *pgdX* (42), and genes of recombination pathways (*ruvA*, *ruvB*, *ruvB*, *recR*, *recC*, *xerC*, and *xerD*) required to repair DNA damaged by oxidative stress (43). Several genes implicated in the lung model (*nadN*, *hel*, and *pdgX*) are dispensable for bloodstream colonization by *H. influenzae* type b (41, 44). It is possible that *H. influenzae* strains differ in their requirements for these genes in vivo, or that these genes are specifically needed in the lung, where nucleotide sources and levels of oxidative stress may differ from those in the blood.

Conclusion

HITS provides a massively parallel system to simultaneously monitor the relative fitness of thousands of individual mutants undergoing a selection condition of interest. In this report, a large library of $\approx$75,000 *H. influenzae* mutants was subjected to selection in a murine pulmonary model of pathogenesis to identify genes required for prolonging survival of *H. influenzae* in the lung. Analysis of the mutant library by HITS was easily performed, highly reproducible, and remarkably comprehensive. Sequencing of transposon/chromosome junctions revealed independent insertions in nearly 56,000 genomic sites. More than 96% of *H. influenzae* protein coding genes were analyzed using a conservative cutoff that excluded 35 genes that were duplicated or contained repetitive sequences and 25 genes that had $<$8 TA dinucleotides available for *mariner* insertion. It is anticipated that with improvements in high-throughput sequencing technology, the depth of sequencing coverage will substantially increase to allow an even greater level of resolution and dynamic range. The results provide a genome-wide assessment of the genetic requirements of this bacterium for growth or survival in the lung, and also represent the most comprehensive fitness analysis that has been applied to *H. influenzae* mutants in any animal model.

The profile of genes required in this environment provides a view of the host–pathogen interactions occurring during pulmonary pathogenesis and will provide insight into potential strategies for the design of vaccines or therapeutics to specifically target *H. influenzae* in this site of disease.

The HITS procedure was demonstrated using a *mariner* transposon bank in *H. influenzae*; however, none of the procedures are organism specific, and the approach should be applicable to any organism amenable to mutational analysis with transposons. A major advantage of the approach we present in this report is that it can be applied to existing mutant libraries and does not require use of a specifically engineered transposon. In fact, the procedure should be readily adaptable to libraries generated with any insertion mutation capable of providing a primer-binding site. Although a complete genome sequence is useful for HITS, mapping insertions to annotated contigs of draft genome sequences should yield much of the same information. Although HITS was used in this report to obtain a genome-wide assessment of the requirements for lung pathogenesis, the procedure should be equally effective for analysis of requirements for growth or survival under any selective condition that can be applied to large populations of mutants *en masse*. Because of the speed and resolution of HITS, it will be possible to efficiently conduct fitness analyses in diverse contexts of host–microbe interactions. Application of this approach is expected to generate multifaceted views of the genetic requirements of pathogens in the environments they encounter in diverse stages of pathogenesis.

Materials and Methods

High-Density Mutagenesis of *H. influenzae* by in Vitro Transposition. *H. influenzae* Rd strain BA042 and clinical isolate nontypeable strain NT127 (21) were grown in brain heart infusion broth (BHI) supplemented with 10 μg/mL hemin and 10 μg/mL NAD (sBHI) or on sBHI agar plates at 35 °C. Media contained kanamycin sulfate at 20 μg/mL (sBHI-Kan) where indicated. The mini-*mariner* transposon *mmTrcK* (carried on plasmid pENTtrcK) was derived from *magellan1* (8) by replacement of the endogenous promoter for the kanamycin resistance gene, *aphI*, with the *trc* promoter. Transposition reactions were performed in vitro as described in ref. 45. Transposition products were transformed into *H. influenzae* as described previously (8, 46). After selection on sBHI-Kan plates, the insertion library (≈75,000 colonies) was harvested in BHI with 20% glycerol and stored at −80 °C.

Selection of Transposon Insertion Mutant Library in the Lung Model. The *H. influenzae* insertion library (3.1×10^{10} cfu) was inoculated in 50 mL sBHI and grown with shaking at 225 rpm to a final OD_{600} of 0.45. For representation of the input library, cells from 35 mL of culture were collected by centrifugation and stored at −80 °C. Inoculum for murine lung infection was prepared by pelleting 5 mL of the culture, washing in 1× Hank's buffered salt solution, and dilution to concentration of 2.5×10^8 cfu/mL. Forty microliters (10^7 cfu) was inoculated into the nares of 5 female C57BL/6 mice (7 to 8 weeks old) anesthetized with ketamine (50 mg/kg) and xylazine (5 mg/kg) by i.p. injection. At 24 h of infection, lungs were harvested and homogenized using a Fisher Tissuemiser. Dilutions of homogenates were plated on sBHI to enumerate total cfu per lung. To recover the output library, homogenates from each mouse lung were plated on 12 sBHI agar plates, and resulting colonies were collected for chromosomal DNA isolation via phenol chloroform extraction (8). All experiments with mice were conducted with prior approval of the University of Massachusetts Institutional Animal Use and Care Committee (IACUC).

Genetic Footprinting. Genetic footprinting was conducted on *H. influenzae* genomic DNA from input and output libraries as described elsewhere (45) with transposon-specific primer, marout, and gene-specific primers that bind 5′ or 3′ of each gene. Primer design, PCR conditions, and image analysis are described in *SI Methods*, and footprinting primers are listed in Table S4.

Illumina Sequencing of Transposon–Chromosome Junctions from Mutant Libraries. Genomic DNA from mutant libraries prepared before and after in vivo selection was sheared using a Covaris S2 device. Paired-end Illumina libraries were created by ligation of adaptors to sheared DNA as described by Bentley et al. (47) and size selected between 200 and 400 bp. Enrichment of transposon/chromosomal junction regions was performed by PCR amplification with a 5′ biotinylated transposon enrichment primer, PE1MAR, and adapter-specific PCR PE2.0 enrichment primer (Table S4). Thermocycler settings were as follows: 30 s, 98 °C; 18 cycles of 10 s, 98 °C, 30 s, 65 °C, 30 s, 72 °C; 5 min, 72 °C. Fragments between 250 and 300 bp were gel purified and added to Dynal MyOne C1 beads (Invitrogen) to capture biotinylated templates containing transposon insertions. The beads were washed according to the manufacturer's instructions, and the nonbiotinylated strand was eluted with 125 mM NaOH. Supernatants were recovered from beads, neutralized, and templates purified with MinElute PCR purification columns (Qiagen). The resulting transposon libraries were quantified on an Agilent Bioanalyzer 2100 RNA Pico6000 chip (Agilent Technologies). Single-stranded templates were cluster amplified and sequenced on an Illumina GAII, as described in ref. 47.

Analysis and Mapping of Illumina Sequencing Data. The Illumina sequencing reads that contained the *Himar1* inverted terminal repeat (ITR) sequence and the adjacent TA insertion site were identified in the raw fasta files and trimmed of the ITR sequence. The processed sequencing reads are provided as multifasta files for Input Library Sample1 (Dataset S1), Input Library Sample2 (Dataset S2), and Lung Output Library (Dataset S3). Processed reads, typically 53 bp in length, were aligned to the *H. influenzae* Rd KW20 genome sequence (48) (GenBank accession no. L42023) using SOAPv1.11 alignment software using default settings (2 mismatches allowed per read) (49). A custom PERL script was used to parse insertion site coordinates from the SOAP output file to report the number of reads mapped per site and strand orientations of aligned reads (*SI Computer Script*). The data were imported into Microsoft Excel, and insertion site coordinates were mapped to positions within protein coding genes annotated in protein table RefSeq file NC_000907.ptt (from the National Center for Biotechnology Information: ftp://ftp.ncbi.nih.gov/). For each gene, the number of insertion sites identified and the total number of sequencing reads in the internal 5–80% of the gene were determined using Excel functions. Additional details are provided in *SI Methods*. A draft version of *H. influenzae* strain RdAW (also referred to as BA042) genome sequence was generated and is at least 99.98% identical to strain Rd KW20 (48).

Single-Strain Infections in the Pulmonary Clearance Model. Nonpolar mutations deleting the *galU* and *orfH* genes were introduced into nontypeable *H. influenzae* strain, NT127 (*SI Methods*). Each *H. influenzae* strain was used to inoculate C57BL/6 mice by the intranasal route as described above. At 24 h of infection, mice were killed and bacterial cfu in the lungs were enumerated as described above. The number of cfu recovered from the lungs of each mouse was compared by one-way ANOVA with Tukey's multiple comparison test. Blood samples obtained immediately before killing revealed no detectable cfu. Procedures were approved by the University of Massachusetts IACUC.

Note Added in Proof. During review of this manuscript we learned of an independent report submitted to *Nature Methods* by Tim van Opijnen, Kip L. Bodi, and Andrew Camilli, in which transposon junction sequencing was successfully applied to study genetic networks (personal communication).

ACKNOWLEDGMENTS. We thank John Leong for his helpful comments; and David Lapointe (UMass Medical School) and David Borenstein, Philip Montgomery, and Carsten Russ (Broad Institute) for analytic support and technical input that contributed to the development of HITS methodology. This project has been funded in part by the National Institute of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN266200400001C (Broad Institute), and by National Institutes of Health Grant 1RO1-AI49437 (to B.J.A.).

1. Hensel M, et al. (1995) Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269:400–403.
2. Sassetti CM, Boyd DH, Rubin EJ (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci USA* 98:12712–12717.
3. Salama NR, Shepherd B, Falkow S (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J Bacteriol* 186:7926–7935.
4. Badarinarayana V, et al. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat Biotechnol* 19:1060–1065.
5. Vila-Corcoles A, et al. (2009) Epidemiology of community-acquired pneumonia in older adults: A population-based study. *Respir Med* 103:309–316.
6. Sethi S, Murphy TF (2001) Bacterial infection in chronic obstructive pulmonary disease in 2000: A state-of-the-art review. *Clin Microbiol Rev* 14:336–363.
7. Lampe DJ, Churchill ME, Robertson HM (1996) A purified *mariner* transposase is sufficient to mediate transposition in vitro. *EMBO J* 15:5470–5479.
8. Akerley BJ, et al. (1998) Systematic identification of essential genes by in vitro *mariner* mutagenesis. *Proc Natl Acad Sci USA* 95:8927–8932.

9. Rubin EJ, et al. (1999) In vivo transposition of *mariner*-based elements in enteric bacteria and mycobacteria. *Proc Natl Acad Sci USA* 96:1645–1650.
10. Wieland CW, et al. (2005) The MyD88-dependent, but not the MyD88-independent, pathway of TLR4 signaling is important in clearing nontypeable *Haemophilus influenzae* from the mouse lung. *J Immunol* 175:6042–6049.
11. Toews GB, Hart DA, Hansen EJ (1985) Effect of systemic immunization on pulmonary clearance of *Haemophilus influenzae* type b. *Infect Immun* 48:343–349.
12. Toews GB, Viroslav S, Hart DA, Hansen EJ (1984) Pulmonary clearance of encapsulated and unencapsulated *Haemophilus influenzae* strains. *Infect Immun* 45:437–442.
13. Singh IR, Crowley RA, Brown PO (1997) High-resolution functional mapping of a cloned gene by genetic footprinting. *Proc Natl Acad Sci USA* 94:1304–1309.
14. Gronow S, Brabetz W, Lindner B, Brade H (2005) OpsX from *Haemophilus influenzae* represents a novel type of heptosyltransferase I in lipopolysaccharide biosynthesis. *J Bacteriol* 187:6242–6247.
15. Nichols WA, et al. (1997) Identification of the ADP-L-glycero-D-manno-heptose-6-epimerase (rfaD) and heptosyltransferase II (rfaF) biosynthesis genes from nontypeable *Haemophilus influenzae* 2019. *Infect Immun* 65:1377–1386.
16. Hood DW, et al. (2001) Genetic basis for expression of the major globotetraose-containing lipopolysaccharide from *H. influenzae* strain Rd (RM118). *Glycobiology* 11:957–967.
17. Nesper J, et al. (2001) Characterization of *Vibrio cholerae* O1 El tor *galU* and *galE* mutants: influence on lipopolysaccharide structure, colonization, and biofilm formation. *Infect Immun* 69:435–445.
18. Weissborn AC, Liu Q, Rumley MK, Kennedy EP (1994) UTP: Alpha-D-glucose-1-phosphate uridylyltransferase of *Escherichia coli*: Isolation and DNA sequence of the *galU* gene and purification of the enzyme. *J Bacteriol* 176:2611–2618.
19. Choudhury B, Carlson RW, Goldberg JB (2005) The structure of the lipopolysaccharide from a galU mutant of *Pseudomonas aeruginosa* serogroup-O11. *Carbohydr Res* 340:2761–2772.
20. Rosadini CV, Wong SM, Akerley BJ (2008) The periplasmic disulfide oxidoreductase DsbA contributes to *Haemophilus influenzae* pathogenesis. *Infect Immun* 76:1498–1508.
21. Harrington JC, et al. (2009) Resistance of *Haemophilus influenzae* to reactive nitrogen donors and gamma interferon-stimulated macrophages requires the formate-dependent nitrite reductase regulator-activated *ytfE* gene. *Infect Immun* 77:1945–1958.
22. Hood DW, et al. (1996) Use of the complete genome sequence information of *Haemophilus influenzae* strain Rd to investigate lipopolysaccharide biosynthesis. *Mol Microbiol* 22:951–965.
23. Wong SM, Alugupalli KR, Ram S, Akerley BJ (2007) The ArcA regulon and oxidative stress resistance in *Haemophilus influenzae*. *Mol Microbiol* 64:1375–1390.
24. Figueira MA, et al. (2007) Role of complement in defense of the middle ear revealed by restoring the virulence of nontypeable *Haemophilus influenzae siaB* mutants. *Infect Immun* 75:325–333.
25. Ho DK, et al. (2007) lgtC expression modulates resistance to C4b deposition on an invasive nontypeable *Haemophilus influenzae*. *J Immunol* 178:1002–1012.
26. Lysenko ES, et al. (2000) Bacterial phosphorylcholine decreases susceptibility to the antimicrobial peptide LL-37/hCAP18 expressed in the upper respiratory tract. *Infect Immun* 68:1664–1671.
27. Hood DW, et al. (2004) Three genes, *lgtF*, *lic2C* and *lpsA*, have a primary role in determining the pattern of oligosaccharide extension from the inner core of *Haemophilus influenzae* LPS. *Microbiology* 150:2089–2097.
28. Pang B, et al. (2008) Lipooligosaccharides containing phosphorylcholine delay pulmonary clearance of nontypeable *Haemophilus influenzae*. *Infect Immun* 76:2037–2043.
29. De Buck E, Lammertyn E, Anne J (2008) The importance of the twin-arginine translocation pathway for bacterial virulence. *Trends Microbiol* 16:442–453.
30. Malinverni JC, Silhavy TJ (2009) An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. *Proc Natl Acad Sci USA* 106:8009–8014.
31. Suzuki T, et al. (1994) Identification and characterization of a chromosomal virulence gene, *vacJ*, required for intercellular spreading of *Shigella flexneri*. *Mol Microbiol* 11:31–41.
32. Hong M, Gleason Y, Wyckoff EE, Payne SM (1998) Identification of two *Shigella flexneri* chromosomal loci involved in intercellular spreading. *Infect Immun* 66:4700–4710.
33. Cuccui J, et al. (2007) Development of signature-tagged mutagenesis in *Burkholderia pseudomallei* to identify genes important in survival and pathogenesis. *Infect Immun* 75:1186–1195.
34. Haseltine WA, Block R (1973) Synthesis of guanosine tetra- and pentaphosphate requires the presence of a codon-specific, uncharged transfer ribonucleic acid in the acceptor site of ribosomes. *Proc Natl Acad Sci USA* 70:1564–1568.
35. Paul BJ, et al. (2004) DksA: A critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell* 118:311–322.
36. Kuroda A, et al. (2001) Role of inorganic polyphosphate in promoting ribosomal protein degradation by the Lon protease in *E. coli*. *Science* 293:705–708.
37. Butler SM, Festa RA, Pearce MJ, Darwin KH (2006) Self-compartmentalized bacterial proteases and pathogenesis. *Mol Microbiol* 60:553–562.
38. Kemmer G, et al. (2001) NadN and e (P4) are essential for utilization of NAD and nicotinamide mononucleotide but not nicotinamide riboside in *Haemophilus influenzae*. *J Bacteriol* 183:3974–3981.
39. Lamarche MG, Wanner BL, Crepin S, Harel J (2008) The phosphate regulon and bacterial virulence: A regulatory network connecting phosphate homeostasis and pathogenesis. *FEMS Microbiol Rev* 32:461–473.
40. Wanner BL (1993) Gene regulation by phosphate in enteric bacteria. *J Cell Biochem* 51:47–54.
41. Vergauwen B, Herbert M, Van Beeumen JJ (2006) Hydrogen peroxide scavenging is not a virulence determinant in the pathogenesis of *Haemophilus influenzae* type b strain Eagan. *BMC Microbiol* 6:3.
42. Harrison A, et al. (2006) The OxyR regulon in nontypeable *Haemophilus influenzae*. *J Bacteriol* 189:1004–1012.
43. Stohl EA, Seifert HS (2006) Neisseria gonorrhoeae DNA recombination and repair enzymes protect against oxidative damage caused by hydrogen peroxide. *J Bacteriol* 188:7645–7651.
44. Schmidt-Brauns J, et al. (2001) Is a NAD pyrophosphatase activity necessary for *Haemophilus influenzae* type b multiplication in the blood stream? *Int J Med Microbiol* 291:219–225.
45. Akerley BJ, Lampe DJ (2002) Analysis of gene function in bacterial pathogens by GAMBIT. *Methods Enzymol* 358:100–108.
46. Barcak GJ, Chandler MS, Redfield RJ, Tomb JF (1991) Genetic systems in *Haemophilus influenzae*. *Methods Enzymol* 204:321–342.
47. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
48. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.
49. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: Short oligonucleotide alignment program. *Bioinformatics* 24:713–714.

Table S2. Genes required for growth or survival in the lung model detected by HITS

HI Locus	% TA Hit[a]	Survival Index[b]	COG[c]	Gene	Description
HI0761	74.6	0.000	M	mltC	murein transglycosylase C
HI0621.1	58.1	0.000	E	gmhB	D,D-heptose 1,7-bisphosphate phosphatase
HI0857	57.1	0.000	S	zapA	cell division associated protein
HI0187a	53.8	0.000	U	tatA	Sec-independent protein secretion pathway component TatA
HI0337	53.3	0.000	E	glnB	nitrogen regulatory protein P-II
HI0261	52.7	0.000	M	opsX	heptosyltransferase I
HI0854	51.1	0.000	P	-	putative heme iron utilization protein
HI1033	51.0	0.000	E	serB	phosphoserine phosphatase
HI1181	48.0	0.000	G	gmhA	phosphoheptose isomerase
HI0836	47.8	0.000	J	genX	lysyl-tRNA synthetase
HI1114	46.5	0.000	M	rfaD	ADP-L-glycero-D-mannoheptose-6-epimerase
HI0309	45.7	0.000	L	xerD	site-specific tyrosine recombinase XerD
HI0706	45.5	0.000	M	nlpD	lipoprotein
HI1086	44.1	0.000	Q	yrbE (mlaE)	ABC transporter permease
HI0032	43.5	0.000	M	pbp2	penicillin-binding protein 2
HI0847	42.9	0.000	S	-	hypothetical protein HI0847
HI0031	41.7	0.000	D	rodA	rod shape-determining protein
HI1105	40.0	0.000	M	rfaF	heptosyltransferase II
HI0812	68.4	0.011	M	galU	UDP-glucose pyrophosphorylase
HI0523	51.1	0.012	M	orfH	heptosyltransferase III
HI0428	46.4	0.012	O	dsbB	disulfide bond formation protein B
HI0942	45.5	0.014	L	recC	exodeoxyribonuclease V gamma chain
HI0551	55.3	0.014	T	apaH	diadenosine tetraphosphatase
HI0740	51.7	0.015	G	yhxB	phosphomannomutase (pgm)
HI0465	67.2	0.017	H	serA	D-3-phosphoglycerate dehydrogenase
HI0846	60.0	0.021	O	por	periplasmic oxidoreductase DsbA
HI1193	42.6	0.021	E	ilvE	branched-chain amino acid aminotransferase
HI0221	49.3	0.022	F	guaB	inositol-5-monophosphate dehydrogenase
HI0693	70.7	0.022	R	hel	lipoprotein E
HI0314	43.5	0.022	L	ruvC	Holliday junction resolvase
HI0141	43.5	0.023	G	nagB	glucosamine-6-phosphate deaminase
HI0334	54.5	0.024	T	relA	GTP pyrophosphokinase
HI0351	50.8	0.025	M	galE	UDP-glucose 4-epimerase
HI0770	65.3	0.028	D	ftsX	cell division protein FtsX
HI1087	54.3	0.028	Q	yrbF (mlaF)	ABC transporter ATPase
HI0290	60.8	0.029	P	-	putative cation-transporting ATPase
HI0286	59.6	0.031	E	alaT	aminotransferase AlaT
HI0039	57.9	0.033	M	mreD	rod shape-determining protein MreD
HI0138	51.9	0.033	L	rnhA	ribonuclease H
HI0312	57.4	0.038	L	ruvB	Holliday junction DNA helicase B
HI0361	43.8	0.040	P	yfeB	iron (chelated) transporter ATP-binding protein
HI0313	64.3	0.043	L	ruvA	Holliday junction DNA helicase motor protein
HI0464	67.7	0.048	G	rpiA	ribose-5-phosphate isomerase A
HI1617	57.1	0.050	E	aspC	aromatic amino acid aminotransferase
HI0066	48.1	0.051	M	amiB	N-acetylmuramoyl-L-alanine amidase
HI0769	67.7	0.052	D	ftsE	cell division ATP-binding protein
HI1702	56.0	0.060	E	metE	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase
HI0896	60.6	0.063	D	ftsN	cell division protein
HI0639	54.4	0.065	F	purB	adenylosuccinate lyase
HI1633	68.9	0.066	F	purA	adenylosuccinate synthetase
HI0479	48.8	0.067	C	atpD	F0F1 ATP synthase subunit beta
HI0206	49.0	0.068	F	nadN	NAD nucleotidase
HI1380[d]	45.7	0.070	P	pstB	phosphate transport system ATPase component

HI1290	40.7	0.070	E	tyrA	bifunctional chorismate mutase/prephenate dehydrogenase
HI0408	51.0	0.080	P	yebM	ABC transporter ATP-binding protein
HI0572	48.5	0.080	O	pgdX	peroxiredoxin hybrid Prx5
HI0461	62.3	0.081	S	-	hypothetical protein HI0461
HI0188	72.3	0.081	U	tatC	Sec-independent protein translocase protein TatC
HI0718	68.4	0.083	M	vacJ (mlaA)	lipoprotein
HI0676	48.8	0.083	L	xerC	site-specific tyrosine recombinase XerC
HI1145	59.4	0.084	E	pheA	chorismate mutase/prephenate dehydratase
HI0209	46.8	0.085	L	dam	DNA adenine methylase
HI1146	59.2	0.086	R	-	hypothetical protein HI1146
HI1277	45.1	0.086	D	mrp	putative ATPase
HI1494	40.0	0.091	V	-	N-acetylmuramoyl-L-alanine amidase
HI1615	61.9	0.096	F	purE	phosphoribosylaminoimidazole carboxylase catalytic subunit
HI1698	51.3	0.099	M	-	lipopolysaccharide biosynthesis protein
HI0407	54.5	0.100	P	-	hypothetical protein HI0407
HI0119	45.8	0.101	P	znuA	high-affinity zinc transporter periplasmic component
HI1263	46.7	0.102	E	metX	homoserine O-acetyltransferase
HI0329	59.5	0.102	E	-	hypothetical protein HI0329
HI0187b	75.0	0.102	U	tatB	sec-independent translocase
HI1248	63.5	0.103	R	-	hypothetical protein HI1248
HI1323	42.9	0.103	S	-	hypothetical protein HI1323
HI0887	63.9	0.105	F	purH	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase
HI1249	71.4	0.106	R	-	hypothetical protein HI1249
HI0480	42.9	0.106	C	atpG	F0F1 ATP synthase subunit gamma
HI0086	68.4	0.109	E	metB	cystathionine gamma-synthase
HI0371	50.0	0.111	S	-	hypothetical protein HI0371
HI0669	40.0	0.111	C	mioC	flavodoxin
HI1389.1	50.7	0.112	E	trpC	bifunctional indole-3-glycerol phosphate synthase/phosphoribosylanthranilate isomerase
HI0330	43.9	0.114	M	oapA	opacity associated protein
HI1388	46.2	0.115	E	trpG	anthranilate synthase component II
HI1077	51.7	0.116	F	pyrG	CTP synthetase
HI1084	61.2	0.116	Q	yrbC (mlaC)	ABC transporter
HI1432	47.5	0.118	E	trpA	tryptophan synthase subunit alpha
HI1389	44.2	0.119	E	trpD	anthranilate phosphoribosyltransferase
HI0122	65.8	0.123	E	metC	cystathionine beta-lyase
HI0055	62.1	0.123	G	uxuA	mannonate dehydratase
HI0950	50.0	0.125	J	rpmG	50S ribosomal protein L33
HI1085	60.0	0.130	Q	yrbD (mlaD)	ABC transporter periplasmic protein
HI0416	52.4	0.133	H	thiD	phosphomethylpyrimidine kinase
HI1383m	54.2	0.141	P	pstS	phosphate ABC transporter phosphate-binding protein
HI1726	52.1	0.148	F	hemH	phosphoribosylaminoimidazole-succinocarboxamide synthase
HI1605	45.2	0.150	T	-	hypothetical protein HI1605
HI1431	47.9	0.151	E	trpB	tryptophan synthase subunit beta
HI1381	54.2	0.152	P	pstA	phosphate ABC transporter permease
HI0653	62.5	0.152	M	lgtF	LOS glycosyltransferase
HI1382	51.5	0.154	P	pstC	phosphate ABC transporter permease
HI1379	59.1	0.158	T	phoB	phosphate regulon transcriptional regulatory protein PhoB
HI0427	60.2	0.162	P	nhaB	sodium/proton antiporter
HI1713	50.0	0.167	G	ptsH	phosphocarrier protein HPr
HI0325	42.2	0.168	R	-	hypothetical protein HI0325
HI0462	52.1	0.172	O	lon	ATP-dependent proteinase
HI0756	62.0	0.177	D	-	hypothetical protein HI0756
HI0888	41.4	0.181	F	purD	phosphoribosylamine--glycine ligase
HI1207	72.7	0.184	F	purF	amidophosphoribosyltransferase
HI0518	62.5	0.185	F	deoD	purine nucleoside phosphorylase
HI0457	50.0	0.188	R	-	hypothetical protein HI0457

HI0752	60.2	0.188	F	purL	phosphoribosylformylglycinamidine synthase
HI0062	52.9	0.190	T	dksA	dnaK suppressor protein
HI1387	49.4	0.195	E	trpE	anthranilate synthase component I
HI0415	41.4	0.213	H	thiM	hydroxyethylthiazole kinase
HI0980	76.2	0.217	K	fis	DNA-binding protein Fis
HI1171	54.5	0.225	E	trpG	para-aminobenzoate synthase component II
HI0176	42.6	0.231	J	rluD	23S rRNA pseudouridine synthase D
HI0038	47.1	0.235	M	mreC	rod shape-determining protein
HI1429	61.4	0.238	F	purM	phosphoribosylaminoimidazole synthetase
HI1658	59.3	0.241	R	-	hypothetical protein HI1658
HI1647	52.4	0.241	H	pdxS	pyridoxal biosynthesis lyase PdxS
HI1642	42.0	0.241	V	sapF	anti peptide resistance ABC transporter ATPase
HI0564	53.2	0.242	E	asnA	asparagine synthetase AsnA
HI1159m	48.7	0.243	O	ybbN	thioredoxin domain-containing protein
HI0535	56.0	0.245	O	ureH	urease accessory protein
HI0443	63.6	0.246	L	recR	recombination protein RecR
HI0011	68.2	0.256	L	holD	DNA polymerase III subunit psi
HI0571	57.8	0.256	K	oxyR	DNA-binding transcriptional regulator OxyR
HI0765	63.9	0.258	M	lpsA	LOS glycosyltransferase
HI0713	66.1	0.270	O	tig	trigger factor
HI0411	91.7	0.274	R	hfq	RNA-binding protein Hfq
HI0029	52.0	0.276	M	dacA	penicillin-binding protein 5
HI1004	59.8	0.279	O	ppiD	peptidyl-prolyl cis-trans isomerase
HI1103	42.5	0.282	E	cysK	cysteine synthetase
HI1191	65.5	0.285	R	-	hypothetical protein HI1191
HI0534	57.3	0.289	E	aspA	aspartate ammonia-lyase
HI1234	73.7	0.290	G	mgsA	methylglyoxal synthase

The set of *H. influenzae* genes listed in the table were determined by two criteria as depicted in Fig. 2. Genes were considered to be specifically required *in vivo* if they sustained insertions in at least 40% of the possible insertion sites in the internal 5-80% of the gene in the *in vitro* grown input library and the total number of sequencing reads mapping to insertion sites in each gene decreased at least 3.3-fold after *in vivo* passage relative to the input library.

[a] Percentage of TA dinucleotides with detected insertions in the 5’ 5 to 80% of each gene.

[b] Survival index calculated as the fraction of total sequencing reads mapped to insertions in the output library relative to the input library.

[c] Identifier for the Cluster of Orthologous Groups functional classification.

[d] For HI1380, the gene *pstB* has been annotated as a containing an “artificial frameshift” and was not listed in the *H. influenzae* Rd KW20 (NC_000907.ptt) protein table.

Table S3. Distribution of genes required *in vivo* organized by functional categories

	Genes required *in vivo*		Total *H. influenzae* genes	
Category[a]	No. of genes	Percent in category	No.of genes[c]	Percent in category
Amino acid transport and metabolism (E)	23[b]	16.9%	148	8.9%
Cell wall/membrane/envelope biogenesis (M)	18[b]	13.2%	118	7.1%
Nucleotide transport and metabolism (F)	13[b]	9.6%	56	3.4%
DNA Replication, recombination and repair (L)	10	7.4%	108	6.5%
Inorganic ion transport and metabolism (P)	10	7.4%	85	5.1%
Posttranslational modification, chaperones (O)	8	5.9%	85	5.1%
Carbohydrate transport and metabolism (G)	7	5.1%	104	6.3%
Cell division and chromosome partitioning (D)	6[b]	4.4%	23	1.4%
Signal transduction mechanisms (T)	5	3.7%	31	1.9%
Coenzyme transport and metabolism (H)	4	2.9%	72	4.3%
2° metabolites biosynthesis and transport (Q)	4[b]	2.9%	14	0.8%
Intracellular trafficking and secretion (U)	3	2.2%	26	1.6%
Translation, ribosomal structure (J)	3	2.2%	148	8.9%
Energy production and conversion (C)	3	2.2%	94	5.7%
Transcription (K)	2	1.5%	73	4.4%
Defense mechanisms (V)	2	1.5%	17	1.0%
Unknown[c]	14	10.3%	404	24.4%
Total	135		1657	

"Percentage in category" refers to the fraction of genes within an individual functional class relative to the total. Letters in parenthesis denote the COG classification identifiers.

[a] Categories in which no genes were identified for growth or survival *in vivo* are not shown

[b] Increases in the number of genes in category identified by HITS relative to the entire genome are significant ($p < 0.05$) by chi squared analysis.

[c] Genes with COG designations of Function Unknown (S), General Function Prediction Only (R) and those without an assigned designation.

Table S4. Oligonucleotides used in this study

Primer name	Sequence (5' to 3')
HITS enrichment primers	
PE1MAR[a,b]	5'-biotinTEG- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGGGACTTATCAGCCAACC
PCR PE2.0[b]	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Genetic Footprinting	
Primer name	Sequence (5' to 3')
marout	CCGGGGACTTATCAGCCAACC
opsX_F	CGCACAATTTACACCAAACCAATCTTCAG
opsX_R	GGCGTATAAGCAAACCTACTTGGATCTCG
rfaF_F	TCATCGTGCCTGTCATCGAATTAAAGTAGG
rfaF_R	CGTAGGGGGATTTGGACCTGTAGTTGTG
orfH_F	GGTAACCCTGAAGGCCCAGATGCAC
orfH_R	GCATCAACGTGGTACGCTCAAGC
galU_F	GGTGGAATGCCAACCGTTCCTTG
galU_R	TGCGGCTTGTGTAAAACGCCCAC
galE_F	ATCACACAACTAATGGACATAGCCCT
galE_R	GATGTTTATCTGTCAGCCAAACAGGCA
xylA_F	CAGTTAATGCACCAACAAGGGTAAGTAAAGCTGAT
xylA_R	CATGTTTTGCGTAACCAACTCCAATCACT
Primers used in mutant construction. Single-strain infections in the pulmonary clearance model	
Primer name	Sequence (5' to 3')
HI0524-int5'	CAAAAAGTTTTTGCTTATGCTAAAG
HI0523out	TTGAAAGTGCGGTCGTATTTTATGT
15kan+HI0523	GATGAGTTTTTCTAAGGAAAAGAGTCTCAAATGAATTTAG
HI0521-int3'	TAAATCGCGTAGCCTTCACATAAATC
HI0523-15	CGACCGCACTTTCAAATGAGCCATATTCAACGGGAAACG
kan3'+TAA	TTAGAAAAACTCATCGAGCATCAAA
812at1945for	GGCGATAAGGTTCGTGTATTAACC
811interg3'2	GTAAAATCCAATGTGATCTCTAAG
764-5'	ATGAATCAGTCAATTTTATC
767-5'	ATGAAAAAAATACAAACGCC

[a] The Illumina primer PCR PE1.0 (underlined) was adapted for enrichment of transposon/chromosomal junctions

[b] Oligonucleotide sequences

Supporting Information

Gawronski et al. 10.1073/pnas.0906627106

SI Text

Analysis and Mapping of Illumina Sequencing Data. The last 20 nucleotides of the PE1MAR enrichment primer are complementary to the inverted terminal repeats of the *himar1*-derived mini-transposon, *mmTrcKan*, and allowed selective enrichment of transposon/chromosomal junctions. The Illumina sequencing primer site was introduced into enriched fragments via PE1MAR immediately 5′ of the transposon-specific sequence. As a result, the sequencing read of an amplified transposon/chromosome junction fragment started with the sequence "cggggacttatcagccaaccTGT*ta*," with the following regions denoted: transposon-specific 3′ end of the PE1MAR primer (lowercase), the remainder of the ITR of the *himar1* transposon (uppercase), and the chromosomal TA insertion site (italics), which is duplicated upon transposon integration. Sequencing reads containing the exact string (above) were trimmed of the ITR-derived sequence to leave the putative TA insertion site at the 5′ end of the sequence. The sequences were aligned to the *H. influenzae* Rd KW20 reference genome (1) using SOAPv1.11 alignment software (2), which allowed for 2 mismatches per read. In the Input Library Sample1 (Dataset S1), 564,484 of the 708,731 trimmed reads (80%) were mapped to the reference genome. In Input Library Sample2 (Dataset S2), 559,459 of the 736,631 reads (76%) and in the Lung Output Library (Dataset S3) 263,237 of the 447,370 (59%) were aligned to the reference genome.

A custom PERL script was used to extract the TA dinucleotide insertion site coordinates from the SOAP output file (*SI Computer Script*). For reads aligning to the plus strand of the genome, the genome coordinate at position 1 of the trimmed read was determined. For reads aligning to the minus strand, the genome coordinate at position 2 of the read was calculated to represent the TA coordinate position with respect to the plus strand. For each TA insertion site detected by alignment, the total number of reads and the strand orientation was determined.

To adjust for sequencing coverage among samples for calculation of survival indices, the number of sequencing reads obtained for each insertion site in the input library was multiplied by a normalization factor of 0.63 defined as (Ro/Ri)/(So/Si), in which the variables represent the total number of sequencing reads (R) and insertion sites (S) detected in the input (i) and output (o) libraries, respectively.

Genetic Footprinting of Input and Lung Selected Libraries. Chromosomal footprinting primers were designed by FastPCR software (www.biocenter.helsinki.fi/bi/programs/fastpcr.htm). The primers were analyzed using standalone MEGAblast (3) against *H. influenzae* Rd KW20 genome (1) (NC_000907.fna) with the options "-W 8 -F F" to search of potential of nonspecific amplification. At secondary sites, the identity at the 3′ end of the primer was <14 bp. PCR reactions consisted of 200 ng library genomic DNA, 1 μM marout primer, 1 μM gene-specific primer, 250 μM dNTP, 2.1 U Taq polymerase, and 0.17 U DeepVent DNA polymerase in 1X Thermopol buffer (NEB) in a 40 μL reaction. The PCR settings were as follows: 95 °C for 2 min; 30 cycles of 94 °C for 30 s, 68 °C for 3.5 min + 10 s per cycle; hold at 8 °C. Primers are listed in Table S4. PCR reactions were examined by agarose gel electrophoresis using 0.9% agarose gels after visualization with ethidium bromide. Gels were imaged using the Kodak Gel Logic 200 imaging system. In the genetic footprinting data, gene positions were determined using the method of Schaffer and Sederoff (4). Migration distances of the 1-kb plus ladder (Invitrogen) were fit using nonlinear least-squares regression by Microsoft Excel and the Solver Add-in.

Single-Strain Infections in the Pulmonary Clearance Model: Strain Construction. The NT127 *orfH* mutant was generated by double-crossover homologous recombination using a kanamycin resistance-marked gene replacement construct. The replacement cassette was assembled from the following 3 fragments: a 1,065-bp PCR product containing the 5′ flanking region of *orfH* was amplified from *H. influenzae* Rd BA042 with primers HI0524-int5′ and HI0523out; a 1,140-bp PCR product containing the 3′ flanking region of *orfH* was amplified from Rd with primers 15kan+HI0523 and HI0521-int3′; and an 815-bp fragment containing the kanamycin resistance gene, *aphI*, from Tn*903* was amplified with primers HI0523–15 and kan3′+TAA (5). Primer sequences are provided in Table S4. Overlap extension PCR of the 3 purified fragments using primers HI0524-int5′ and HI0521-int3′ yielded a final 3.0-kb "stitched" product. The synthetic construct was used to transform *H. influenzae* NT127. Transformants were selected on sBHI containing 20 μg/mL kanamycin and verified by PCR analysis.

The *galU* mutant was generated by transformation of *H. influenzae* NT127 with a PCR product amplified from strain RdgalU (6) using primers 812at1945for and 811interg3′2 (Table S4). The ≈3.3-kb PCR product contains a nonpolar *galU* mutation, in which the *galU* ORF was replaced with that of *aphI* kanamycin resistance gene. Transformants were selected on sBHI agar containing 20 μg/mL kanamycin and verified by PCR analysis.

The *lpsA* mutant strain ΔrepRcp5 was generated by transformation of *H. influenzae* Rd Δrep (7) with a PCR product amplified from *H. influenzae* Rd strain Rcp5 (5) using primers 764–5′ and 767–5′ (Table S4). The strain Δrep contains a deletion of 16 tandem CAAT repeats in *licA* and stabilizes the ORF to prevent stochastic phase variation (5). The ≈4-kb PCR product from Rcp5 contains a transposon insertion mutation in *lpsA* at position 635 of the protein coding sequence (7). Transformants were selected on sBHI agar containing 20 μg/mL kanamycin and verified by PCR analysis. For each strain, 40 μL (10^7 cfu) was inoculated into the nares of 7 C57BL/6 mice (5–8 weeks old) anesthetized with ketamine (50 mg/kg) and xylazine (5 mg/kg) by i.p. injection. At 24 h of infection, lungs were harvested and homogenized using a Fisher Tissuemiser. Dilutions of homogenates were plated on sBHI to enumerate total cfu per lung. Procedures were approved by the University of Massachusetts IACUC.

1. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.
2. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: Short oligonucleotide alignment program. *Bioinformatics* 24:713–714.
3. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7:203–214.
4. Schaffer HE, Sederoff RR (1981) Improved estimation of DNA fragment lengths from agarose gels. *Anal Biochem* 115:113–122.
5. Wong SM, Alugupalli KR, Ram S, Akerley BJ (2007) The ArcA regulon and oxidative stress resistance in *Haemophilus influenzae*. *Mol Microbiol* 64:1375–1390.
6. Rosadini CV, Wong SM, Akerley BJ (2008) The periplasmic disulfide oxidoreductase DsbA contributes to *Haemophilus influenzae* pathogenesis. *Infect Immun* 76:1498–1508.
7. Wong SM, Akerley BJ (2005) Environmental and genetic regulation of the phosphorylcholine epitope of *Haemophilus influenzae* lipooligosaccharide. *Mol Microbiol* 55:724–738.

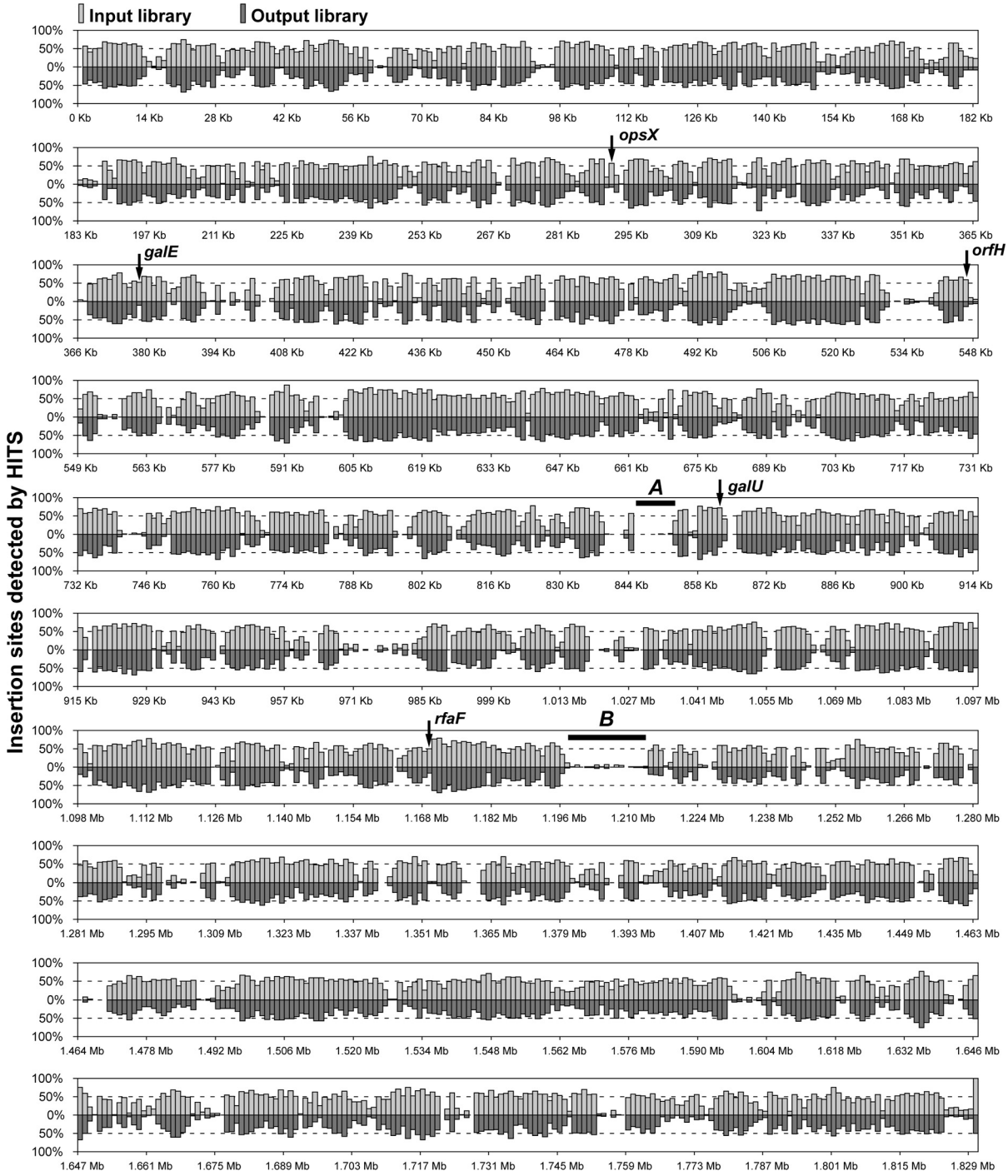

Fig. S1. Genomic distribution of transposon insertions sites in input and output libraries. The percentage of TA dinucleotide sites sustaining insertions identified by HITS analysis are displayed in 1-kb increments along the *H. influenzae* Rd KW20 genome. Light gray bars, input library; dark gray bars, lung-selected output library. Arrows denote the location of several genes discussed in text. Genes located in regions A and B are involved in essential cell functions. Region A: transcription, translation, and protein secretion (pos 846,000–853,000 bp; *rplE*, *rpsN*, *rpsH*, *rplF*, *rplR*, *rpsE*, *rpmD*, *rplO*, *secY*, *rpmJ*, *rpsM*, *rpS11*, *rpsD*, and *rpoA*). Region B: cell division (pos 1,198,000–1,215,000 bp; *ftsI*, *murE*, *murF*, *mraY*, *murD*, *ftsW*, *murG*, *murC*, *ddl*, *ftsQ*, *ftsA*, *ftsZ*, and *lpxC*).

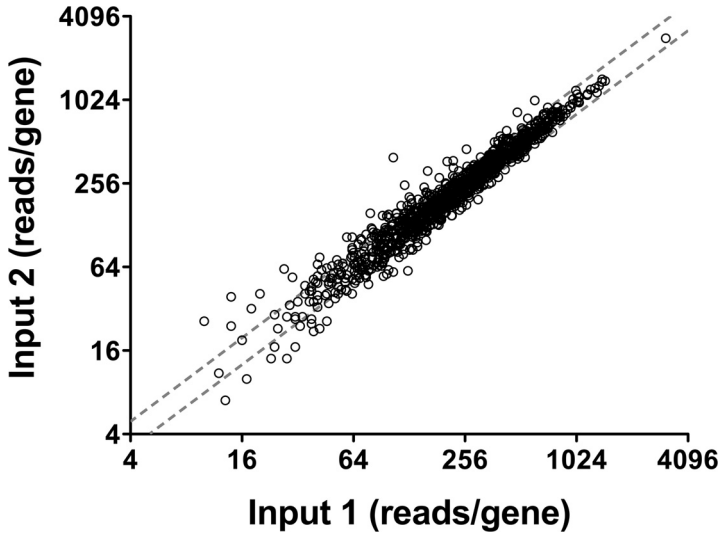

Fig. S2. Replicate analysis of the input library by HITS. Two chromosomal preparations of the input library were analyzed by HITS. Axes represent the total number of sequencing reads mapped to insertion sites in all genes (5′ 5–80% of the coding sequence) for each sample. Dotted lines indicate the upper and lower 20% deviation in x- and y-values from the best fit line ($y = 0.9889x$) determined by linear regression analysis. Of 1,038 genes containing insertions in $>$40% of possible sites, 850 (82%) fell within these boundaries.

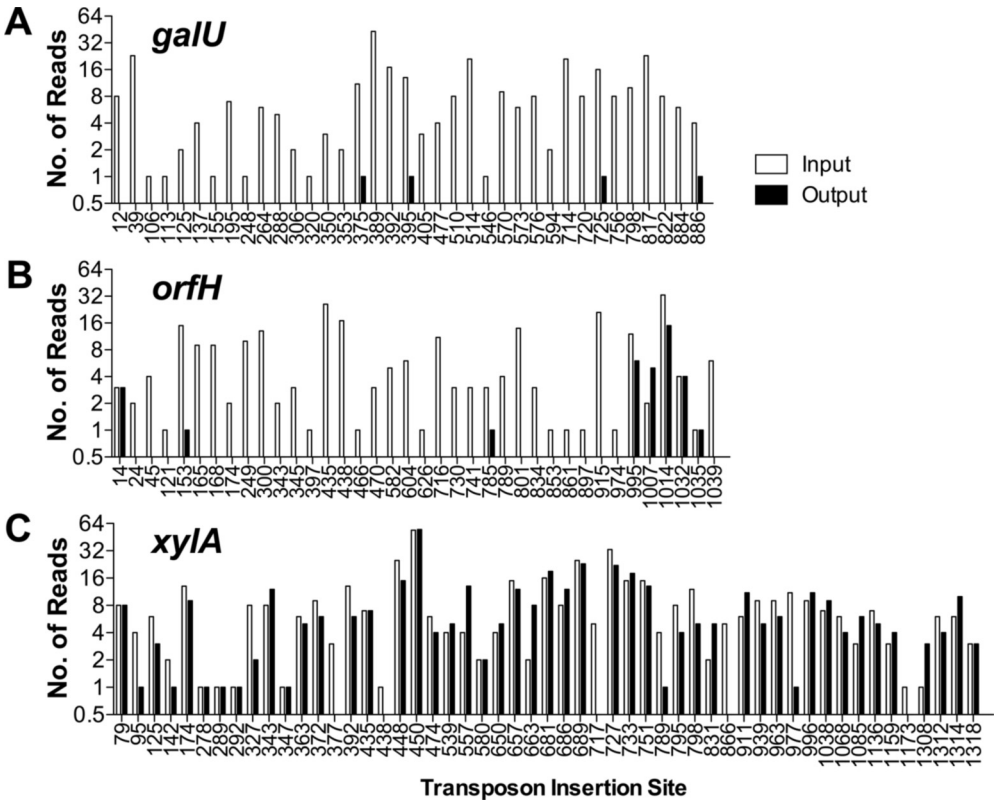

Fig. S3. Insertion patterns in representative genes before and after in vivo selection. Comparison of the number of HITS reads mapped to insertion sites in the input and output pools for *galU* (*A*) *orfH* (*B*), and *xylA* (*C*). The positions of the sites sustaining insertions in the entire coding sequence of the genes are denoted sequentially on the *x* axis, and the numbers of reads mapped to insertions are shown on the *y* axis. Open bars, input library; solid bars, output library. Gene lengths: *galU*, 888 bp: *orfH*, 1,041 bp; *xylA*, 1,320 bp.

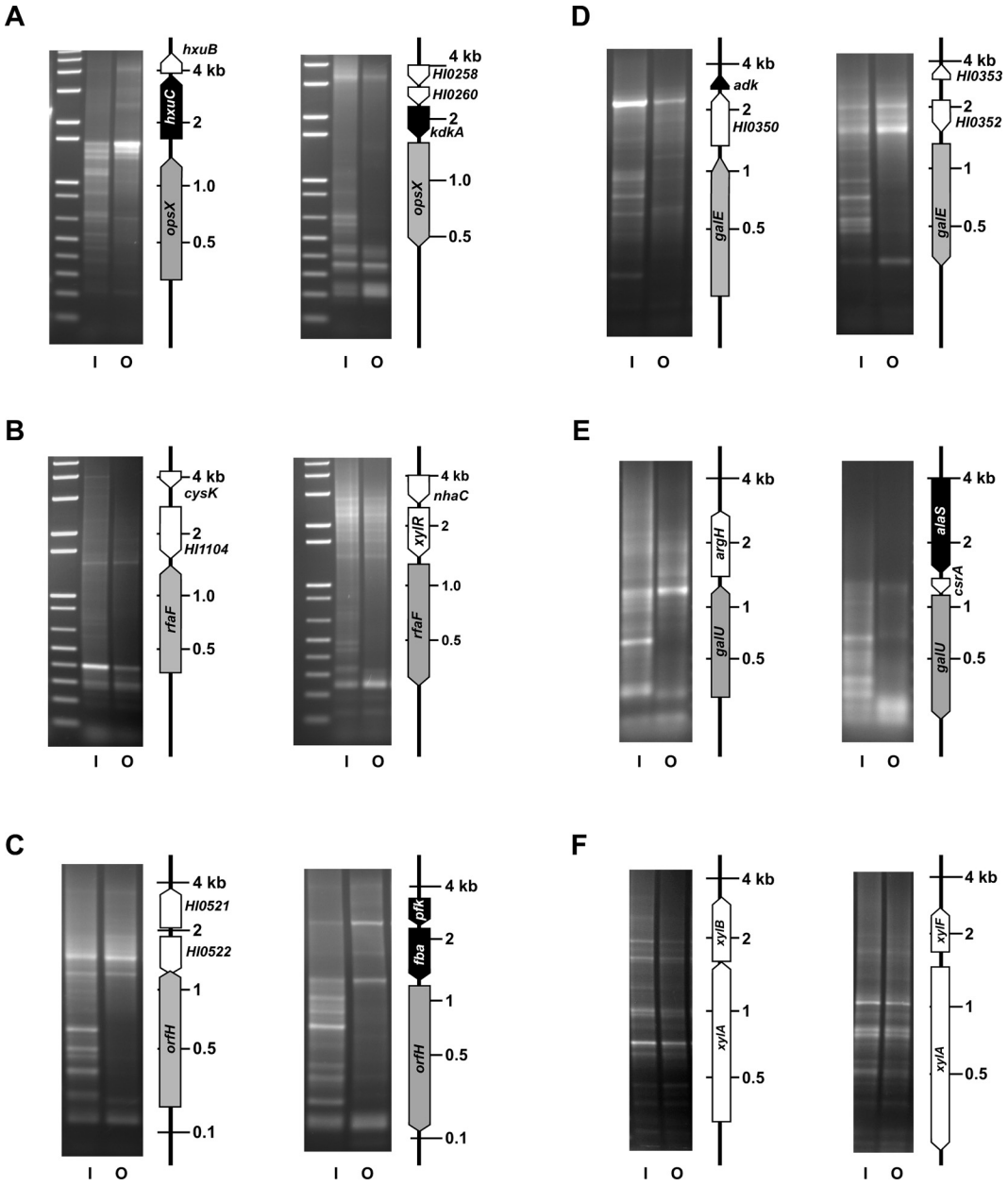

Fig. S4. Genetic footprinting of *H. influenzae* whole-genome transposon insertion library. In each panel *A–F*, the left image shows the genetic footprint analyzed in the forward direction and the right image from the reverse direction. PCR was conducted on genomic DNA from each input and output library using transposon-specific primer, marout, and a chromosomal-specific primer to examine insertions either the 5′ or 3′ direction. I, input library; O, output library. Gene and molecular weight standards positions are displayed to the right of the gel images for each genetic footprint. White, nonessential genes; gray, genes required for growth or survival in the lung; black, genes implicated as essential for growth or survival in vitro. (*A*) Insertion profiles were examined in *opsX*, encoding heptosyltranferase I, using (*Left*) primer opsX_F (positioned 259 bp 5′ of *opsX*) and (*Right*) primer opsX_R (positioned 454 bp 3′ of *opsX*). Similarly, insertions were examined in the following genes using the 5′ primer (*Left*) and the 3′ primer (*Right*): (*B*) rfaF, heptosyltransferase II, rfaF_F (355 bp) and rfaF_R (259 bp); (*C*) *orfH*, heptosyltransferase III, orfH_F (202 bp) and orfH_R (125 bp); (*D*) *galE*, UDP-glucose 4-epimerase, galE_F (162 bp) and galE_R (271 bp); (*E*) *galU*, UDP-glucose pyrophosphorylase, galU_F (278 bp) and galU_R (170 bp); and (*F*) *xylA*, xylose isomerase, xylA_F (279 bp) and xylA_R (153 bp). The primer sequences are provided in Table S4.

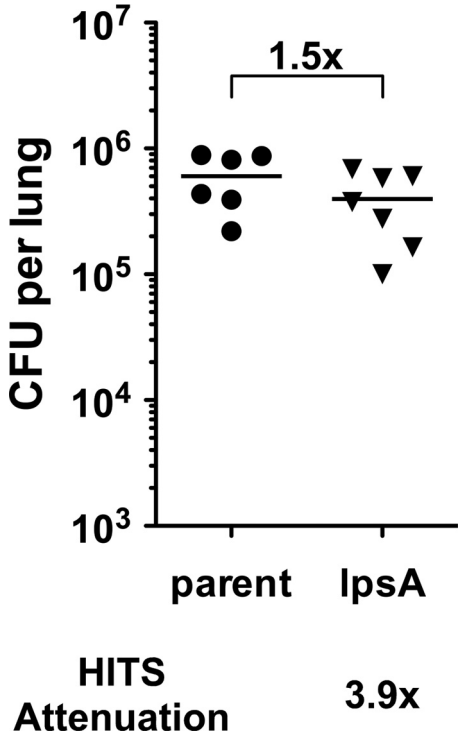

Fig. S5. Effect of *lpsA* mutation in single-strain infections compared with HITS. *H. influenzae* Rd parental strain and the *lpsA* insertion mutant were recovered from lungs of C57BL/6 mice 24 h after intranasal inoculation with each strain. Bars represent the mean cfu per lung (parent, 6.04×10^5 cfu per lung; *lpsA*, 3.98×10^5 cfu per lung). The fold decrease in mean cfu recovered from the *lpsA* mutant vs. Rd parental strain is indicated (bracket). HITS attenuation for *lpsA* was calculated as the reciprocal of the HITS s.i. value of 0.258 and represents the number of reads mapping to *lpsA* in the input library relative to the lung-passaged output library.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Dataset S1
Dataset S2
Dataset S3
SI Computer Script